Pipelines de procesamiento de datos categóricos

Hemos definido como serian los elementos de estadística en el procesamiento de datos, ya hemos hablado del caso con datos numéricos. Pero ¿qué sucede cuando tenemos variables categóricas? El tratamiento ha de ser completamente distinto. En nuestro caso los 2 métodos que existen para procesar variables categóricas son el **Dummy** y **One hot**, de manera que sean fácilmente interpretables por los modelos de aprendizaje automático.

Mapeos de variables categóricas

Dummy:

Es la representación mas compacta que nosotros podamos tener. Mejor uso es cuando mis variables son linealmente independientes y ya lo se a priori. Este concepto tiene que ver con el grado de que no tienen un grado de correlación significativo. Es decir cuando sabemos que las categorías son independientes entre si, de nuestro data set.

One hot

Es la representación mas extensa que podemos tener, es decir, es la representación mas completa para las variables categóricas y la diferencia que tiene respecto al anterior es que permite incluir categorías que no estaban en el data set anteriormente. También el **one hot** lo podemos representar numéricamente.

Ambos siguen la idea de mapear numéricamente, supongamos que tenemos categorías como;**Tipo de motor**=[hibrido,gas,diesel,eléctrico], estas categorías representan algo que es **no ordinal**, es decir; que no tiene un orden.

Entonces aquí las representaciones mas sencillas no sirven porque el algoritmo esta intentando pensar que pueden existir interpolaciones entre una cosa y otra; ejemplo: si yo mapeo que la categoría de **gasolina** es **1**, es decir cada que aparezca la categoría de **gasolina** ponga un **1**, cada vez que aparezca **diesel** ponga un **2**, entonces como el modelo entiende que puede poner números como variables float, entonces el Modelo va interpolar: Si me da un número **1.5** eso ¿qué categoría sería?, esto seria incorrecto y el modelo estaría intentando hacer interpolaciones por naturaleza misma del algoritmo, entonces eso no tiene una interpretación en términos reales de los datos que nosotros estamos usando.

Gracias a este comportamiento la gente de **Data** eligió usar representaciones categóricas tipo **One hot**

Extra:

What is a data pipeline